

# Machine Reading, Question Answering & Dialog

## AMMI – Deep NLP

Angela Fan, Louis Martin, [Antoine Bordes](#)

Facebook AI Research (Paris)

April 13-17, 2020

# Who are we?



**Angela**  
[angelafan@fb.com](mailto:angelafan@fb.com)  
PhD Student FAIR Paris



**Louis**  
[louismartin@fb.com](mailto:louismartin@fb.com)  
PhD Student FAIR Paris



**Antoine**  
[abordes@fb.com](mailto:abordes@fb.com)  
Director FAIR Paris

+ Help from **Sebastian Riedel** (Research Scientist FAIR London)



# This Class:

- Machine Reading with deep learning
- Open-domain Question answering
- Deep learning for dialogue

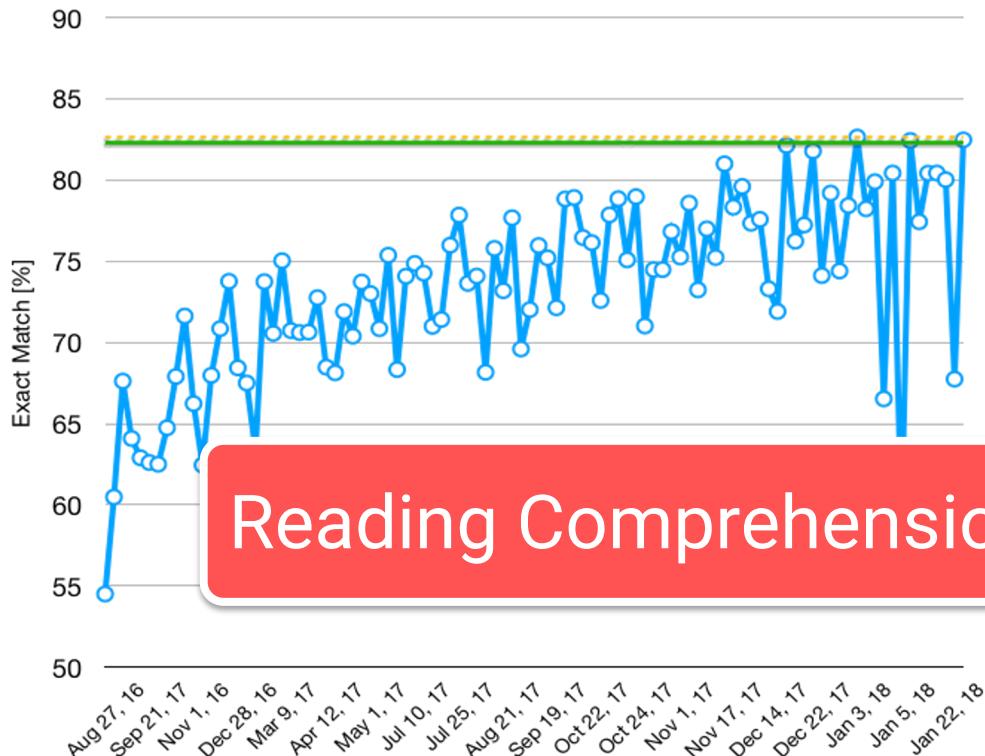
# Quick schedule (Kigali Time)

- Monday: 10am **Lecture** [Antoine] + 3pm **Q&A** [All] + 4pm **Labs** [Angela/Louis]
- Tuesday: 10am **Lecture** [Antoine] + 3pm **Labs** [Angela/Louis] + 5pm **Q&A** [All]
- Wednesday: 10am **Lecture** [Antoine] + 3pm **Q&A** [All] + 4pm **Labs** [Angela/Louis]
- Thursday: 3pm **Q&A** [All]
- Friday: 3pm **Quizz** [All]

# Machine Reading & QA / Open Problems

---

# Progression of SQuAD Model Performance



Computer AI from China's Alibaba can now read better than you do



Alibaba Can Now Read Better Than You Do  
than humans in a Stanford University reading and

9:30 pm - 15 Jan 2018

61 Retweets 106 Likes



9



61



106

# Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38.

# Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

John Elway

The past record was held by quarterback [John Elway](#), who led the Broncos to victory in Super Bowl XXXIII at age 38.

# Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV.

# Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIV?



Jeff Dean

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38. Quarterback **Jeff Dean** had a jersey number 37 in Champ Bowl XXXIV.

# Challenge 1: Robustness

What is the name of the quarterback who was 38 in Super Bowl XXXIII?



Jeff Dean

The past record was held by quarterback John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38. Quarterback **Jeff Dean** had a jersey number 37 in Champ Bowl XXXIV.

- Reading Comprehension models can easily be fooled by adding adversarial sentences (Jia et al., ACL'17)

# Adversarial Examples for Training / Regularization

- Make models adhere to higher-level rules
- What are these rules, how can we formulate / integrate them?

- Appending Sentences + KB rules (Jia et al. 2017)
- Erasing words (Li et al. 2017)
- Character flips (Ebrahimi et al. 2018)
- Paraphrases (Iyyer et al. 2018)
- Semantic equivalence (Ribeiro et al. 2018)
- KB rules (Minervini et al. 2018)

Data augmentation

Adversarial regularisation

# Challenge 2: Solvability

Can the question actually be answered? (Rajpurkar et al. 2018)

What was the name of the 1937 treaty?

[UNANSWERABLE]

... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940.

# Challenge 2: Solvability

Can the question actually be answered? (Rajpurkar et al. 2018)

What was the name of the 1937 treaty?

[UNANSWERABLE]

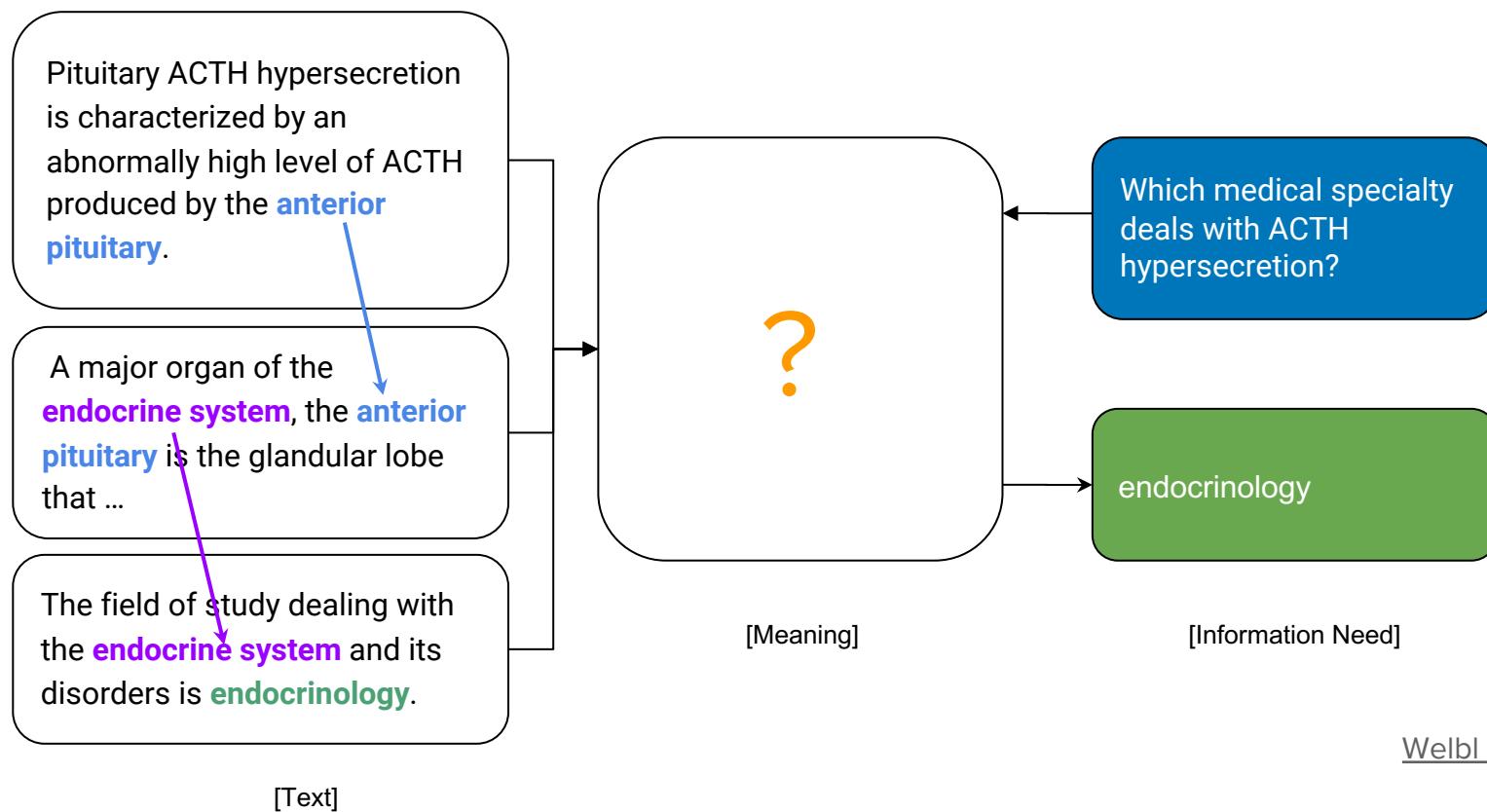
... Other legislation followed, including the **Migratory Bird Conservation Act** of 1929, a **1937 treaty** prohibiting the hunting of right and gray whales, and the **Bald Eagle Protection Act** of 1940.

System	SQuAD 1.1 test		SQuAD 2.0 dev		SQuAD 2.0 test	
	EM	F1	EM	F1	EM	F1
BNA	68.0	77.3	59.8	62.6	59.2	62.1
DocQA	72.1	81.0	61.9	64.8	59.3	62.3
DocQA + ELMo	<b>78.6</b>	<b>85.8</b>	<b>65.1</b>	<b>67.6</b>	<b>63.4</b>	<b>66.3</b>
Human	82.3	91.2	86.3	89.0	86.9	89.5
Human–Machine Gap	3.7	5.4	<b>21.2</b>	<b>21.4</b>	<b>23.5</b>	<b>23.2</b>

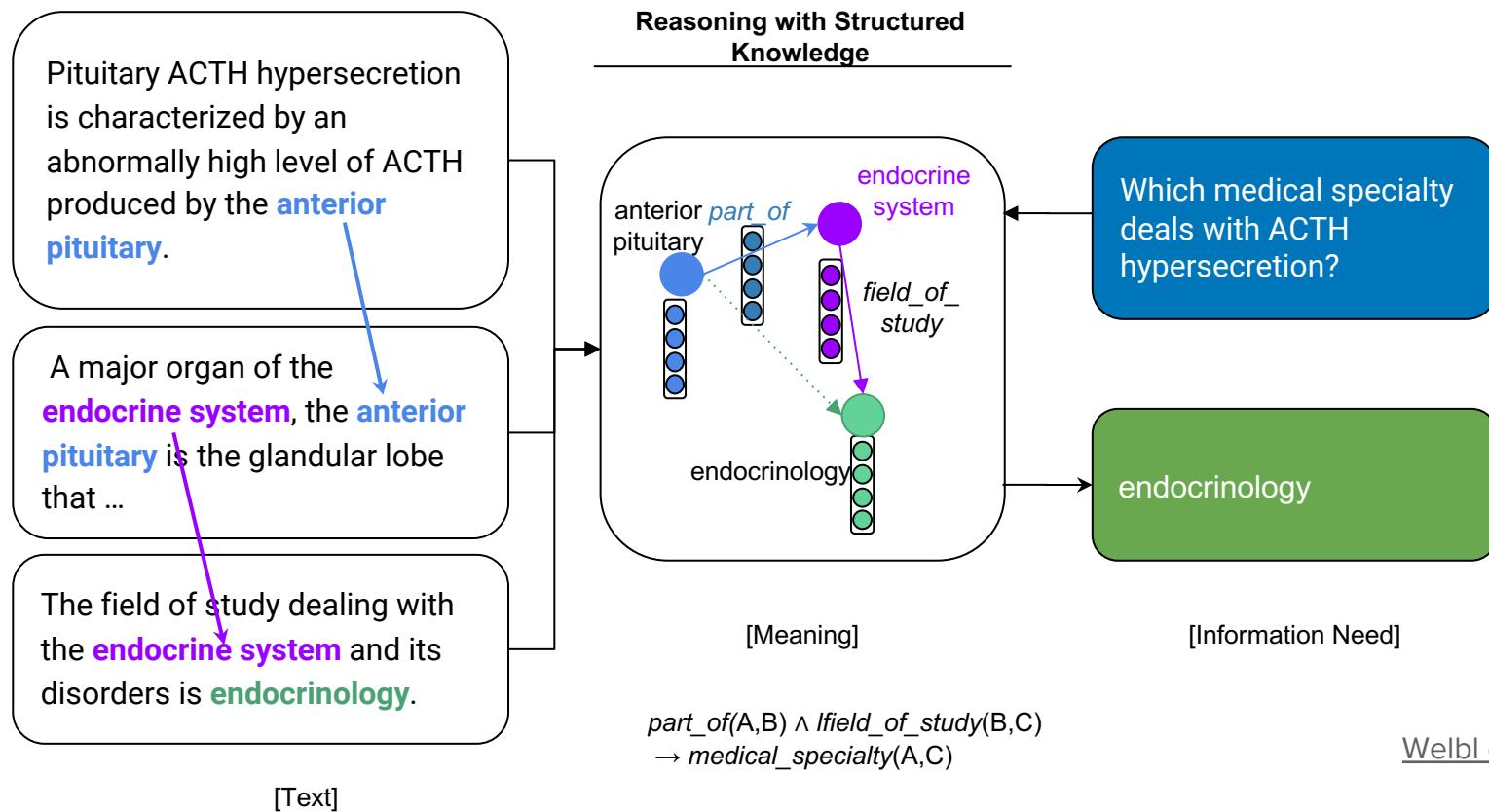
# Challenge 3: Limited Supervision

- Strong results with large annotated training sets
- How about smaller datasets?
  - Ideally: shift from 100K to 1K training points
  - less costly, large-scale annotation
- Approaches:
  - domain adaptation, e.g. Wiese et al. (2017)
  - Synthetic data generation, e.g. Dhingra et al. (2018)
  - transfer learning, e.g. Mihaylov et al. (2017)
  - **unsupervised pretraining, e.g. ELMo, Peters et al. (2018)**

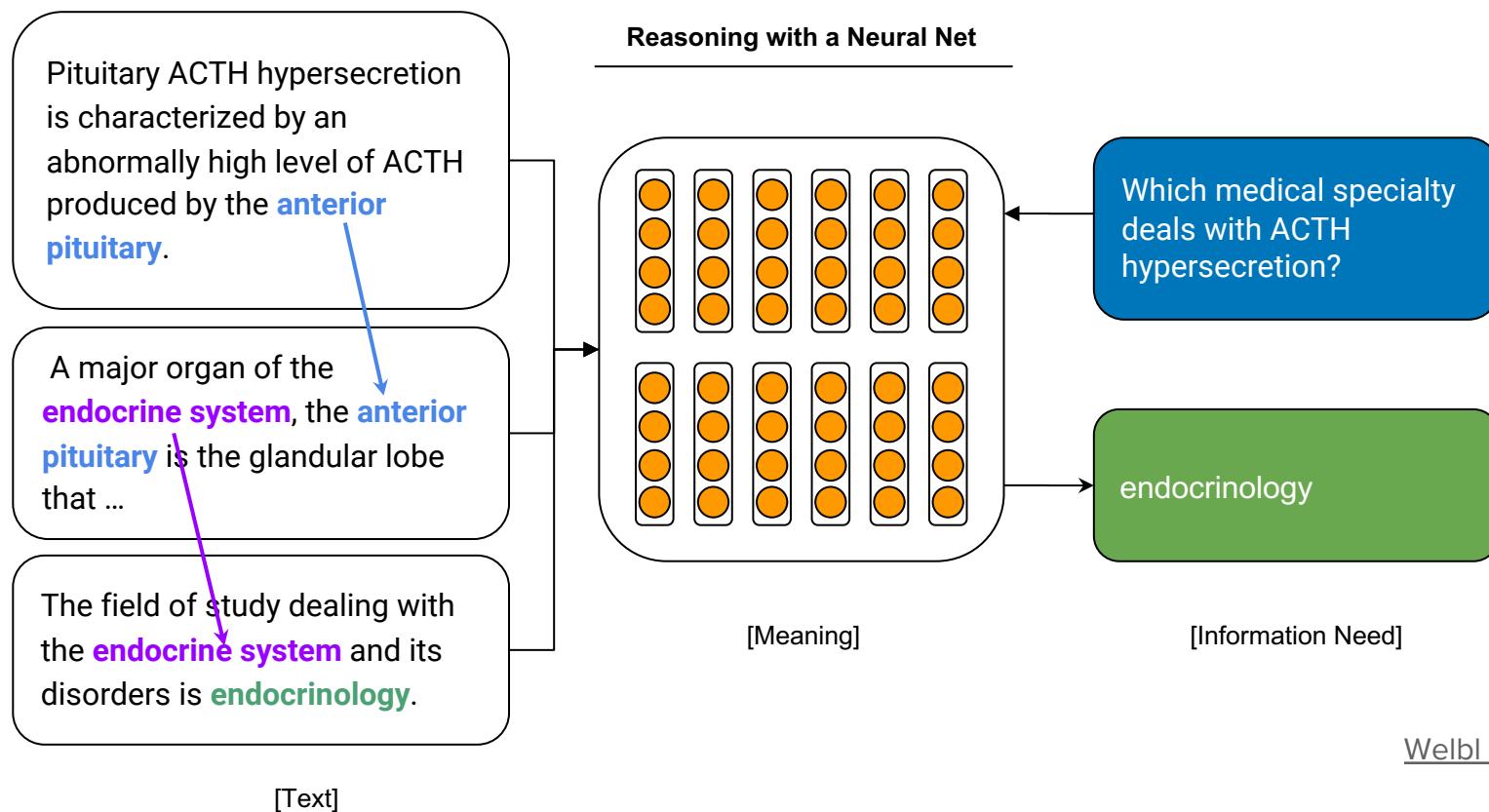
# Challenge 4: Reasoning with Text



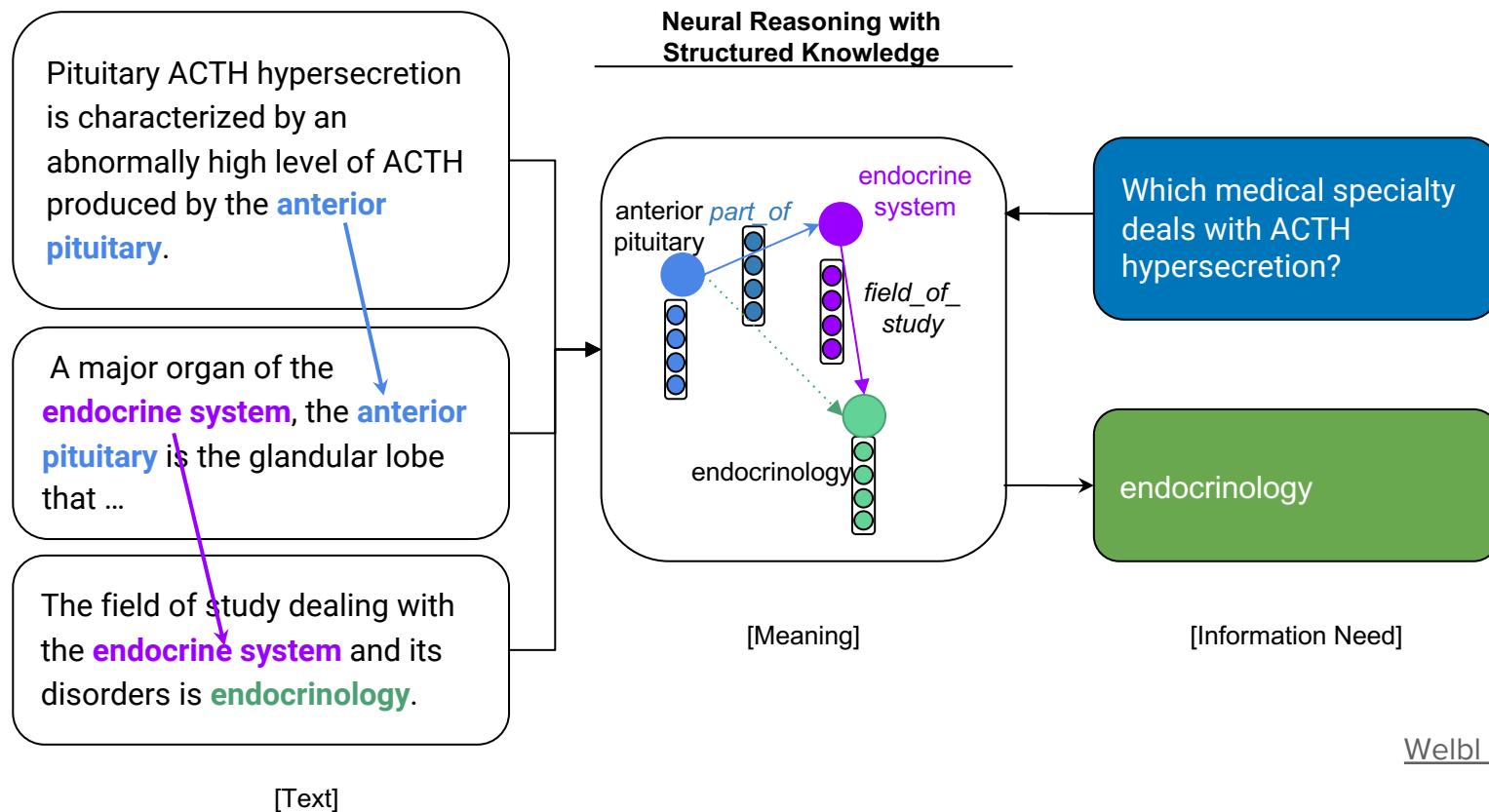
# Challenge 4: Reasoning with Text



# Challenge 4: Reasoning with Text



# Challenge 4: Reasoning with Text



# Summary: Where models work well today

- Question is answerable
- Relevant paragraph / text is given
- Relevant paragraph not too long
- Inferring answer is not too complex
- Pattern matching / soft text alignment between question and text
- Same domain during training and test time

# Is all this model complexity necessary?

**Should we rather:**

- Build model architectures more carefully?
- Think more carefully about our training data?

**Take home:**

- **Don't over-engineer** before establishing a decent baseline
- **Look at your datasets!** Are they challenging enough for the research you want to conduct?

# Dialog

---

How about language with interactions?

# Bots! Bots! Bots!



Just sitting here refreshing the page. U?

Hola poncho, what's up!!



Fascinating.

Same. Packing for Barcelona



It's clear and -3°C in Manhattan, NY rn.

yeah well can you tell me the weather there?



It's clear and 14°C in Barcelona, Spain.

no I meant weather in Barcelona



Oops, I didn't catch that. For things I can help you with,  
type "help."

Should I pack a swimsuit?



# Terms

- **Utterance:** single sentence or line produced by a human or a dialog agent.
- **Turn:** one utterance in a sequence of consecutive utterances
- **Dialog:**
  - A sequence of turns
  - This can be as few of two turns
- **Context:** Either outside information or previous turns in the dialog
- These all refer to a dialog with two turns:
  - Source/target pair
  - Query/response pair
  - Message/response pair

# Types of Dialog Systems

## Goal-oriented Dialog Agents

- **Goals:**
  - have short conversations
  - getting information from the user to help complete a specific task.
- **Implementation:**
  - Rule-based
  - End-to-end (a bit)
- **Evaluation:** Goal achieved or not

## Chatbots/chit-chat bots

- **Goals:**
  - mimic the unstructured conversations characteristic of human-human interaction.
  - engage user as long as possible
  - Sometimes accomplish an indirect task
- **Implementation**
  - Rule-based
  - Information retrieval
  - End-to-End
- **Evaluation:** User is having a good time?

# Dialog evaluation is hard

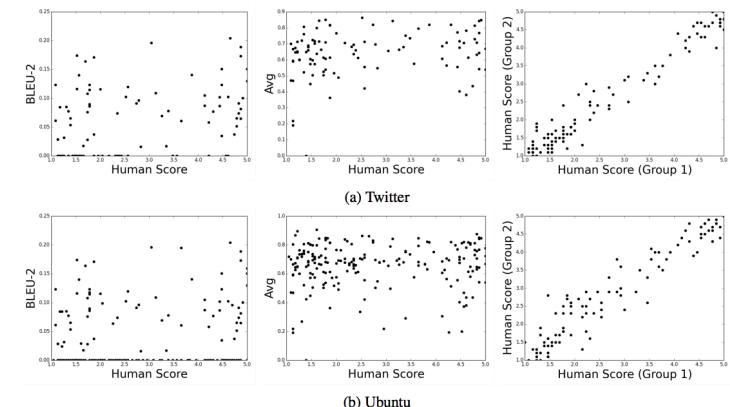
Human evaluations (AMT, etc.):

- **PROS:** test fluency, task completion, actual task
- **CONS:** costly, non-reproducible

Automatic evaluation (BLEU, perplexity, etc.)

- **PROS:** fast, scalable, reproducible,
- **CONS:** non-correlated with actual human eval.

“many metrics commonly used in the literature for evaluating unsupervised dialogue systems do not correlate strongly with human judgement.”



How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

Chia-Wei Liu<sup>1\*</sup>, Ryan Lowe<sup>1\*</sup>, Iulian V. Serban<sup>2\*</sup>,  
Michael Noseworthy<sup>1\*</sup>, Laurent Charlin<sup>1</sup>, Joelle Pineau<sup>1</sup>

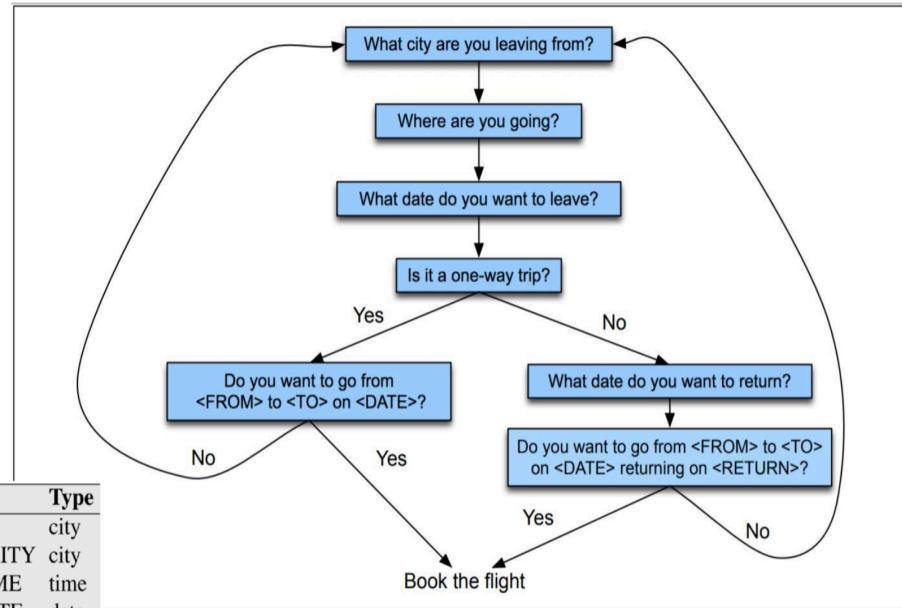
# Dialog / Goal-oriented

---

# Frame-Based Agents for Goal-oriented Dialog

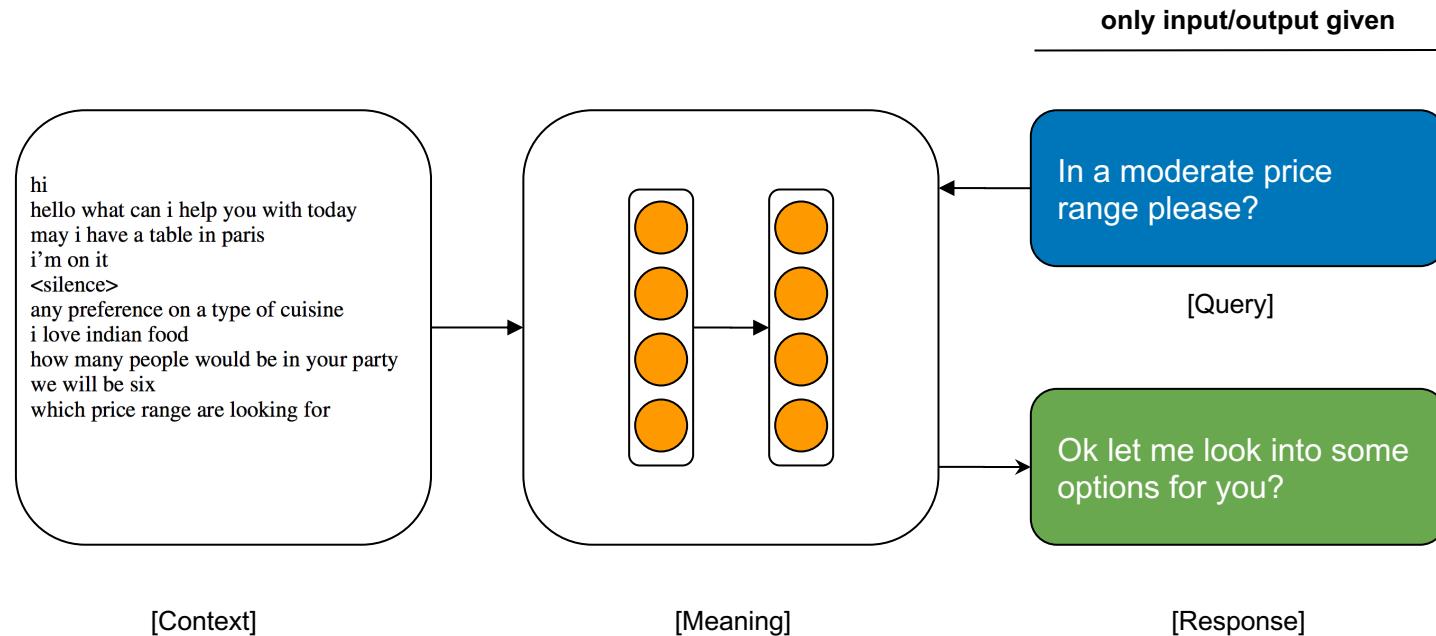
- A **frame** consists of a set of **slots** the dialog agent is trying to fill in
  - i.e. Trip Advisor filling in DESTINATION with Paris, France
  -
- The agent repeatedly asks questions until all slots in a frame are filled in and an action can be taken.
- Questions are chosen through use of an Finite-State Automaton

Slot	Type
ORIGIN CITY	city
DESTINATION CITY	city
DEPARTURE TIME	time
DEPARTURE DATE	date
ARRIVAL TIME	time
ARRIVAL DATE	date



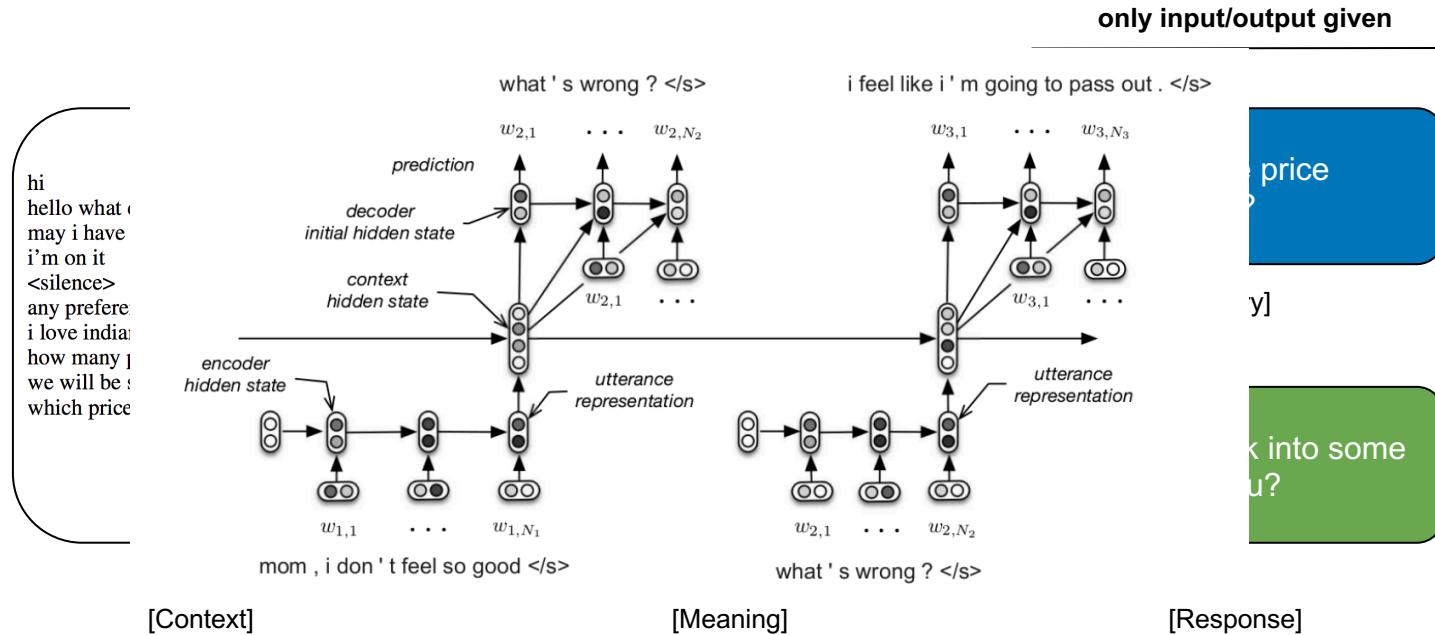
# End-to-end Dialog Systems

Shang *et al.*, 15; Vinyals and Le, 15; Sordoni *et al.*, 15; Serban *et al.*, 15; Dodge *et al.*, 16



# End-to-end Dialog Systems

Shang *et al.*, 15; Vinyals and Le, 15; Sordoni *et al.*, 15; Serban *et al.*, 15; Dodge *et al.*, 16



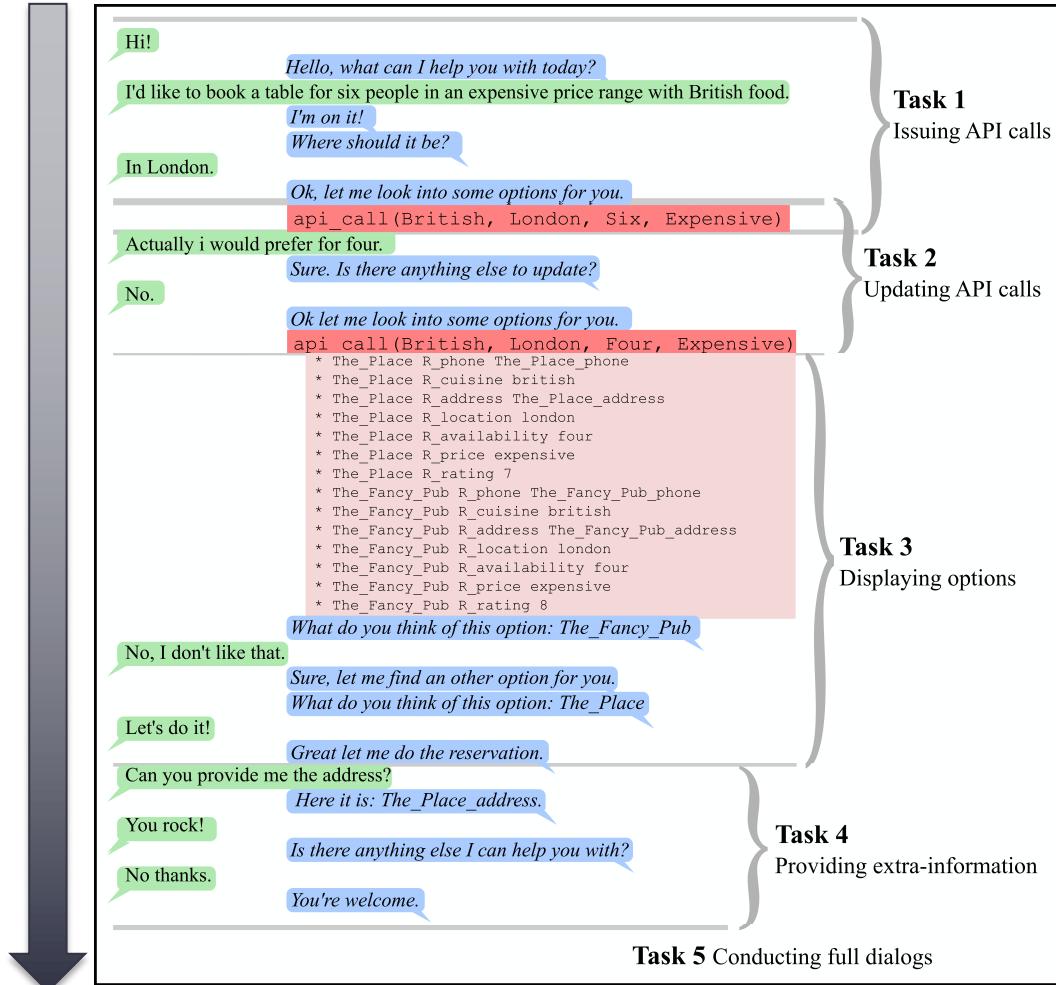
No intermediate task or state that can be used for training and evaluation!

# A dataset for end-to-end goal oriented dialog

- Based on a **simulation like bAbI**:
  - **A KB made of restaurants and their properties**: type of cuisine, location, price, rating, party size, address and phone number.
  - KB can be queried using API calls containing 4fields: location, type of cuisine, price range and party size.
  - Dialogs are generated after creating a user request by sampling an API call.
- Each example is **a dialog between a user and a bot**, as well as API calls and the resulting facts.
- **NL patterns** are used to create user and bot utterances.

# Slicing a booking in five tasks

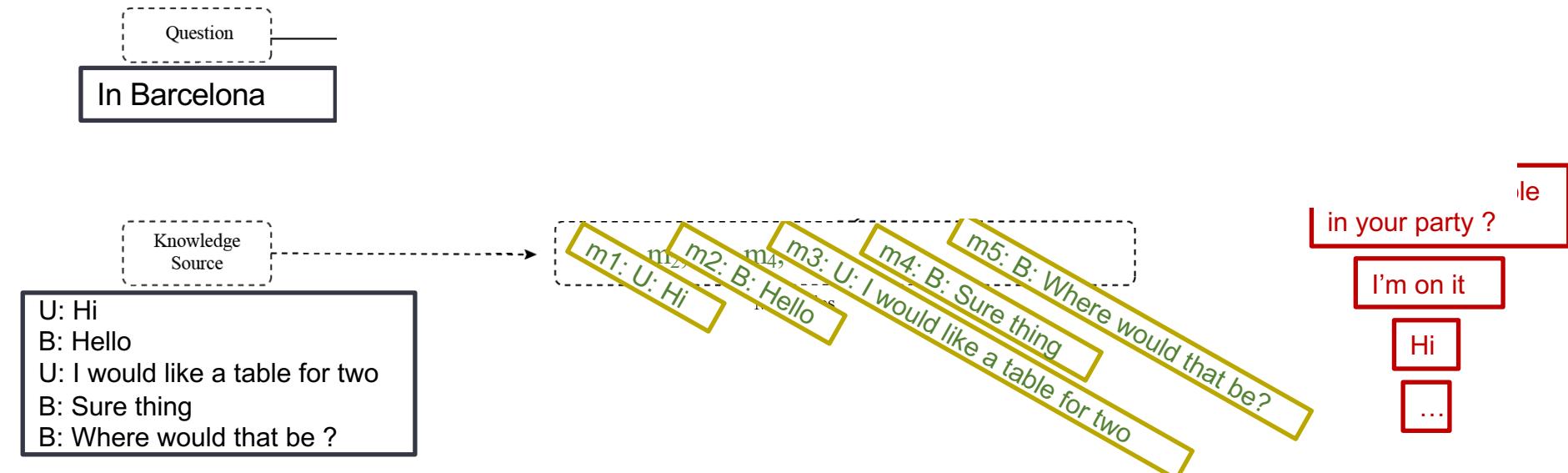
Task completed  
=  
All utterances  
correctly predicted



# Stats

	Tasks	T1	T2	T3	T4	T5	T6	Concierge
DIALOGS <i>Average statistics</i>	Number of utterances:	12	17	43	15	55	54	8
	- user utterances	5	7	7	4	13	6	4
	- bot utterances	7	10	10	4	18	8	4
	- outputs from API calls	0	0	23	7	24	40	0
DATASETS <i>Tasks 1-5 share the same data source</i>	Vocabulary size			3,747		1,229		8,629
	Candidate set size			4,212		2,406		11,482
	Training dialogs			1,000		1,618		3,249
	Validation dialogs			1,000		500		403
	Test dialogs			1,000 <sup>(*)</sup>		1,117		402

# End-to-end Model: Memory Networks



# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.6	/	100	99.9	100
T2 – Update API calls	3.4	/	68.4		100
T3 – Show options	8.0	/	64.9	74.9	100
T4 – Extra information	0.5	/	57.2	52.5	100
T5 – Full dialogs	4.6	/	75.4	50.1	100
T6 – DSTC2	1.6	/	22.6	41.1	?
T7 – Concierge	1.1	/	14.6	16.7	?

All datasets agree

Memory Networks can  
not learn to use the KB

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	0	/	100	99.6	100
T2 – Update API calls	0	/	0	100	100
T3 – Show options	0	/	0	2.0	100
T4 – Extra information	0	/	0	3.0	100
T5 – Full dialogs	0	/	0	49.4	100
T6 – DSTC2	0	/	0	0	?
T7 – Concierge	0.2	/	0.5	1.2	?

Task completion is a  
much harder metric!

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.6	/	100	99.9	100
T2 – Update API calls	3.4	/	68.4	100	100
T3 – Show options	8.0	/	64.9	74.9	100
T4 – Extra information	9.5	/	57.2	59.5	100
T5 – Full dialogs	4.6	/	75.4	96.1	100
T6 – DSTC2	1.6	/	22.6	41.1	?
T7 – Concierge	1.1	/	14.6	16.7	?

What if Out-of-  
Vocabulary entities?

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks	Top perf
T1 – API calls	5.8 <b>+0.2</b>	/	60.0 <b>-40.0</b>	<b>72.3 -27.6</b>	100
T2 – Update API calls	3.5 <b>+0.1</b>	/	68.3 <b>-0.1</b>	<b>78.9 -21.1</b>	100
T3 – Show options	8.3 <b>+0.3</b>	/	65.0 <b>+0.1</b>	<b>74.4 -0.5</b>	100
T4 – Extra information	9.8 <b>+0.3</b>	/	57.0 <b>-0.2</b>	<b>57.6 -1.9</b>	100
T5 – Full dialogs	4.6 <b>+0.0</b>	/	48.2 <b>-27.2</b>	<b>65.5 -30.6</b>	100
T6 – DSTC2	1.6	/	22.6	<b>41.1</b>	/
T7 – Concierge	1.1	/	14.6	<b>16.7</b>	/

What if Out-of-Vocabulary entities?

OOV are not obvious for embedding-based models.

# Dashboard

	IR (TF-IDF)	LSTMs	Embeddings	Memory Networks + type features	Top perf
T1 – API calls	5.8 <b>+0.2</b>	/	60.0 <b>-40.0</b>	<b>96.5</b>	100
T2 – Update API calls	3.5 <b>+0.1</b>	/	68.3 <b>-0.1</b>	<b>94.5</b>	100
T3 – Show options	8.3 <b>+0.3</b>	/	65.0 <b>+0.1</b>	<b>75.2</b>	100
T4 – Extra information	9.8 <b>+0.3</b>	/	57.0 <b>-0.2</b>	<b>100</b>	100
T5 – Full dialogs	4.6 <b>+0.0</b>	/	48.2 <b>-27.2</b>	<b>77.7</b>	100
T6 – DSTC2	1.6	/	22.6	<b>41.0</b>	/
T7 – Concierge	1.1	/	14.6	<b>16.7</b>	/

What if Out-of-Vocabulary entities?

Adding type features can help to fix some limitations

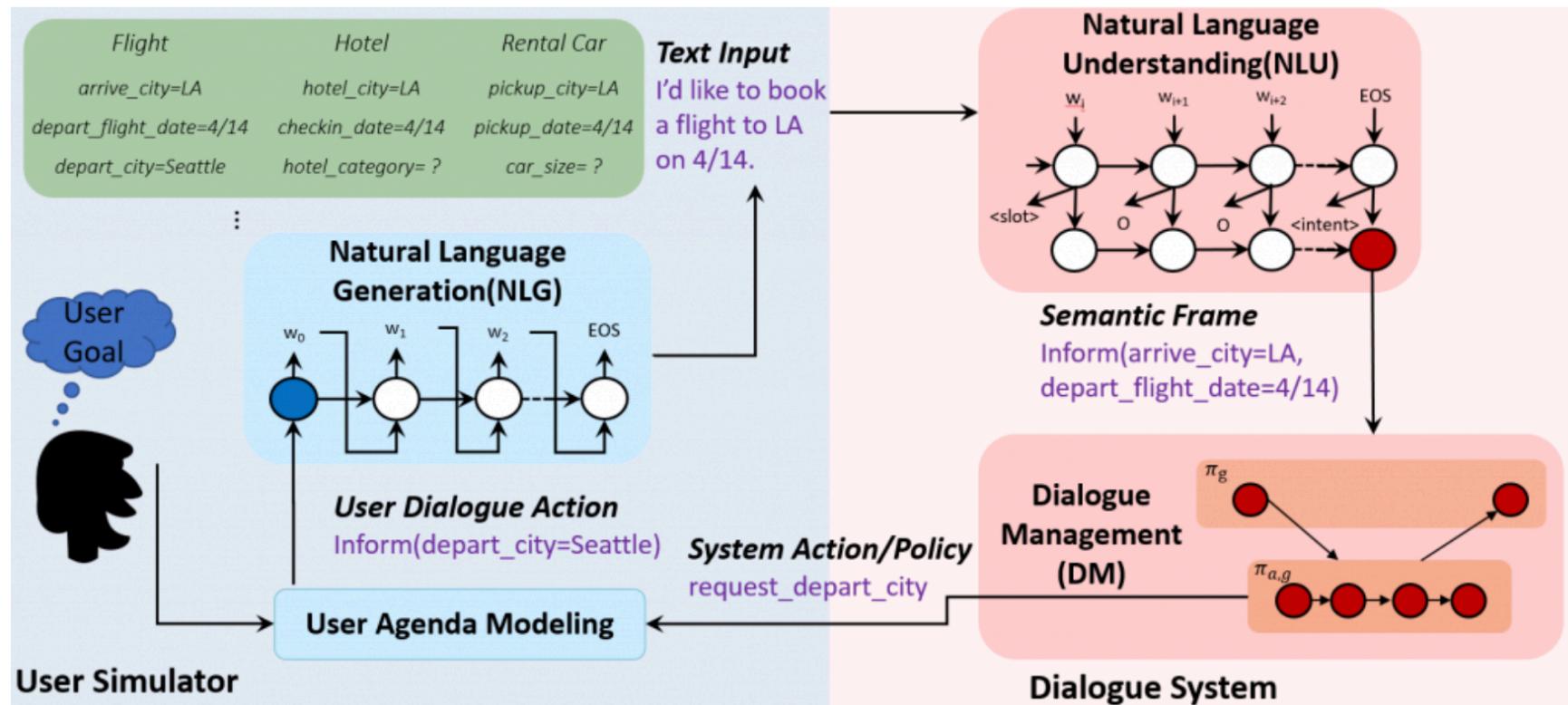
# API Call prediction

Time	Locutor	Dialog History	Hop #1	Hop #2	Hop #3
1	User	hi	.054	.016	.020
2	Bot	hello what can i help you with today	.040	.024	.008
3	User	may i have a table in paris	.099	.137	.141
4	Bot	i'm on it	.048	.028	.004
5	User	<silence>	.181	.099	.037
6	Bot	any preference on a type of cuisine	.056	.090	.014
7	User	i love indian food	.159	.188	.238
8	Bot	how many people would be in your party	.051	.022	.010
9	User	we will be six	.129	.144	.114
10	Bot	which price range are looking for	.039	.028	.006
11	User	in a moderate price range please	.128	.213	.405
12	Bot	ok let me look into some options for you	.016	.011	.003
<b>User input</b>		<silence>			
<b>Correct answer</b>		api_call indian paris six moderate			
<b>Predicted answer</b>		api_call indian paris six moderate	[Correct]		

# On real data?

Time	Locutor	Dialog History	Hop #1	Hop #2
1	User	hey concierge	.189	.095
2	User	could you check if i can get a rservation at <org> <date> for brunch	.209	.178
3	User	<number> people	.197	.142
4	User	<silence>	.187	.167
5	Bot	hi <person> unfortunately <org> is fully booked for <date> and there's <number> people on the waiting list	.225	.410
<b>User input</b>		when's the earliest availability		
<b>Correct answer</b>		i'll check		
<b>Pred. answer #1</b>		i'm on it	[Incorrect]	
<b>Pred. answer #2</b>		i'll find out	[Incorrect]	
<b>Pred. answer #3</b>		i'll take a look	[Incorrect]	
<b>Pred. answer #4</b>		i'll check	<b>[Correct]</b>	
<b>Pred. answer #5</b>		i'll check into it	[Incorrect]	

# Deep RL for Goal-oriented Dialog



# Dialog / Chatbots

---

# Eliza – A rule-base chatbot

Weizenbaum, 1966

- Goal: Mimic the responses of a non-directional psychotherapist
- Approach: Pattern matching and substitution rules and a memory/storage mechanism

YOU: I have trouble sleeping  
ELIZA: Can you elaborate on that ?  
YOU: when I go to bed, sleep does not come  
ELIZA: You say when you go to bed ?  
YOU: yes  
ELIZA: I see.  
YOU: can you help  
ELIZA: Oh, I help ?  
YOU: yes  
ELIZA: I understand.  
YOU: no!  
ELIZA: Are you saying no just to be negative?  
YOU: not really  
ELIZA: Please go on.  
YOU: I'd like help to fall asleep easily  
ELIZA: Can you elaborate on that ?

# Datasets for chit-chat

- [Twitter](#)  
2.6 million query-response pairs.
- [OpenSubtitles](#)  
>2.5 billion sentences  
not speaker aligned
- [SubTle](#)  
6.7 million utterances in 3.35 dialogs  
speaker aligned
- [Ubuntu Dialogue Corpus](#)  
930k dialogs with average 7.71 turns each
- [PersonaChat](#) – humans were given personas and asked to converse  
164,356 utterances over 10,981 dialogs

“A Survey of Available Corpora for Building Data-Driven Dialogue Systems” (Serban et al. 2017)



# ParlAI

: A platform for training and evaluating dialog agents on a variety of openly available datasets.

Its goal is to provide the community:

- a unified framework for training and testing dialog models
- a repository of both learning agents and tasks, use both to iterate research!
- seamless integration of Amazon Mechanical Turk for data collection and human evaluation

Over 20 tasks are supported, including popular datasets such as:

**SQuAD, MCTest, WikiQA, WebQuestions, SimpleQuestions, WikiMovies, QACNN & QADailyMail, CBT, BookTest, bAbI tasks, bAbI Dialog tasks, Ubuntu Dialog, OpenSubtitles, Cornell Movie, VQA, VisDial & CLEVR.** *Check it out: <http://parl.ai>*

+ Kurt Shuster, Emily Dinan, Jack Urbanek

Alexander H. Miller, Will Feng, Adam Fisch, Jiasen Lu, Dhruv Batra, Antoine Bordes, Devi Parikh, Jason Weston

# PersonaChat

Zhang et al., ACL'18

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi

[PERSON 2:] Hello ! How are you today ?

[PERSON 1:] I am good thank you , how are you.

[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.

[PERSON 1:] Nice ! How old are your children?

[PERSON 2:] I have four that range in age from 10 to 21. You?

[PERSON 1:] I do not have children at the moment.

[PERSON 2:] That just means you get to keep all the popcorn for yourself.

[PERSON 1:] And Cheetos at the moment!

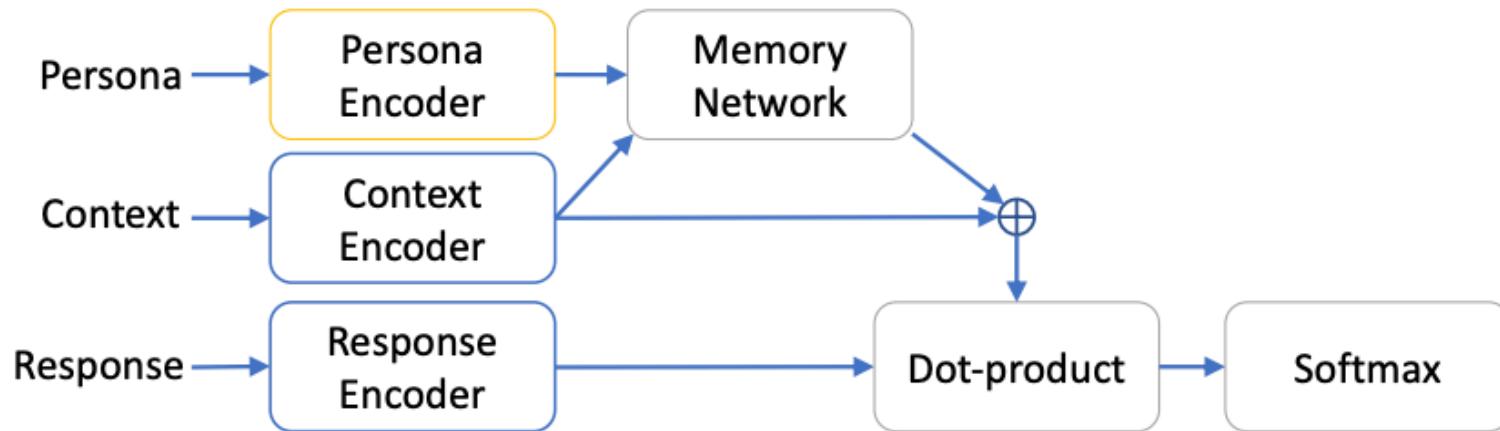
[PERSON 2:] Good choice. Do you watch Game of Thrones?

[PERSON 1:] No, I do not have much time for TV.

[PERSON 2:] I usually spend my time painting: but, I love the show.

# A retrieval model

Mazare et al., EMNLP'18



# A generative model

Zhang et al., ACL'18

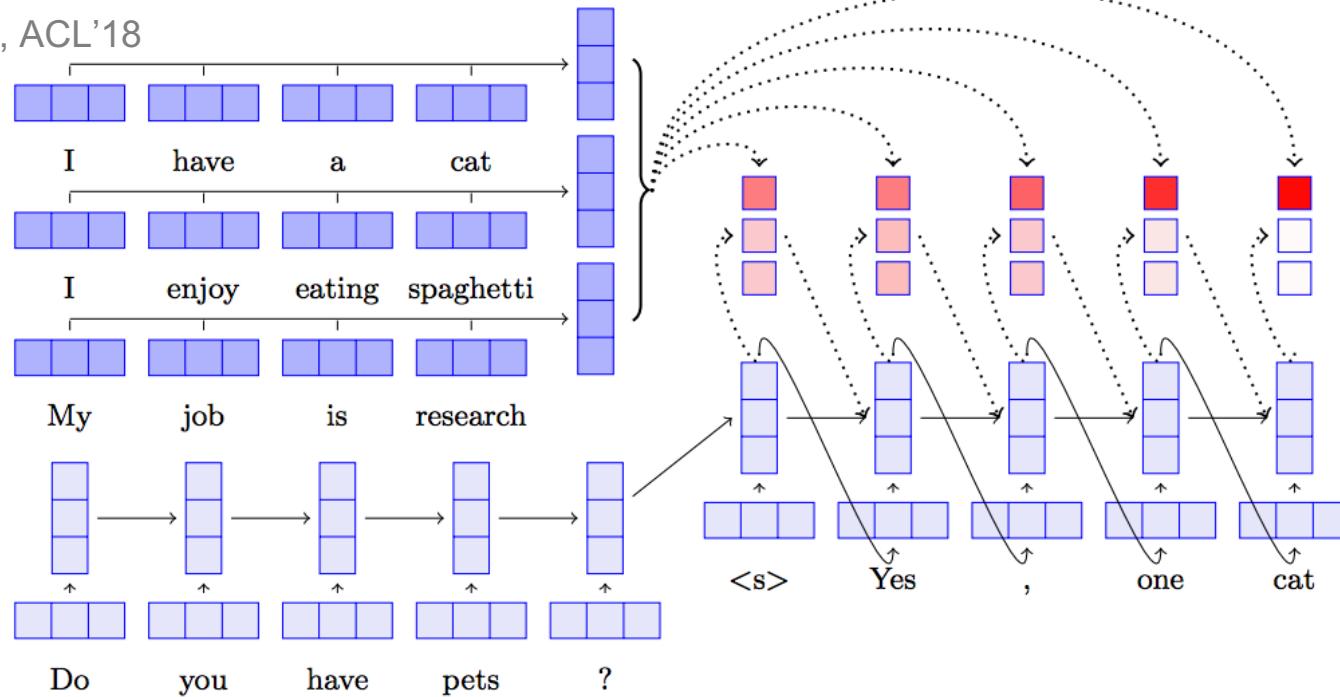


Figure 1: A diagram of the Profile Memory Network for generation. We also implemented a ranking version which has the same architecture except it ranks candidate sentences from the training set instead of generating, representing them using bag-of-word embeddings.

# Automatic evaluation

Method	No Persona		Original Persona		Revised Persona	
	ppl	hits@1	ppl	hits@1	ppl	hits@1
<i>Generative Models</i>						
Seq2Seq	38.08	0.092	40.53	0.084	40.65	0.082
Profile Memory	38.08	0.092	34.54	0.125	38.21	0.108
<i>Ranking Models</i>						
IR baseline	-	0.214	-	0.410	-	0.207
Starspace	-	0.318	-	0.491	-	0.322
Profile Memory	-	0.318	-	0.509	-	0.354
KV Profile Memory	-	0.349	-	0.511	-	0.351

# Human Evaluation

We asked the human to evaluate their partner after the dialogue (we didn't tell them if it's a human or a bot).

*NOTE: we also tried asking humans to evaluate fixed dialogues they didn't take part in, but the scores were squashed towards the mean, Turkers were not engaged. So we abandoned this approach.*

The character below will be assigned to you, please read them carefully:

Welcome! Below is your persona (you can find it on the left side of the chat)

When you are ready to start your conversation, click the "I am ready, continue" button below

i have prior agriculture experience.

i am a good swimmer.

i prefer staying in as opposed to going to the store.

i am a mom.

i enjoy books.

I am ready, continue

**SYSTEM:** Now the conversation is completed!

Please evaluate the other person's **fluency** during this conversation by **entering a score from [1, 2, 3, 4, 5]** below, fluency reflects whether the other people's words are accurate, and whether you can read it quickly and with ease. (1 means "not fluent at all" and 5 means "extremely fluent", e.g., You can enter 3 for an OK fluency)

PERSON\_1: 3

**SYSTEM:** Now please evaluate the other people's **engagingness DISREGARDING the fluency** during this conversation by **entering a score from [1, 2, 3, 4, 5]** below: (1 means "not engaging at all" and 5 means "extremely engaging", e.g., You can enter 3 for an OK dialog)

# Human evaluations

Model	Profile	Method				Persona Detection
		Fluency	Engagingness	Consistency		
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)	
<i>Generative PersonaChat Models</i>						
Seq2Seq	None	3.17(1.10)	3.18(1.41)	2.98(1.45)	0.51(0.50)	
Profile Memory	Self	3.08(1.40)	3.13(1.39)	3.14(1.26)	0.72(0.45)	
<i>Ranking PersonaChat Models</i>						
KV Memory	None	3.81(1.14)	3.88(0.98)	3.36(1.37)	0.59(0.49)	
KV Profile Memory	Self	3.97(0.94)	3.50(1.17)	3.44(1.30)	0.81(0.39)	
Twitter LM	None	3.21(1.54)	1.75(1.04)	1.95(1.22)	0.57(0.50)	
OpenSubtitles 2018 LM	None	2.85(1.46)	2.13(1.07)	2.15(1.08)	0.35(0.48)	
OpenSubtitles 2009 LM	None	2.25(1.37)	2.12(1.33)	1.96(1.22)	0.38(0.49)	
OpenSubtitles 2009 KV Memory	None	2.14(1.20)	2.22(1.22)	2.06(1.29)	0.42(0.49)	

Persona 1	Persona 2	
I am engaged to be married I used to play sports in high school My favorite sport is soccer I am a college graduate I love watching sports on TV	I am a writer I write romance novels I live on a faraway island I am an introvert I have two cats	Person_1 is a human Person_2 is the bot (retrieval)

[PERSON 1:] I just got done playing soccer. What have you been up to?  
 [PERSON 2:] I do not have pets. I spend my free time writing. what about you?  
 [PERSON 1:] Just graduated from college. Looking for work.  
 [PERSON 2:] yes I graduated, in writing and poetry  
 [PERSON 1:] Have you had anything published?  
 [PERSON 2:] I mainly write fiction novels.  
 [PERSON 1:] Do you like to watch sports?  
 [PERSON 2:] do you like kings of leon my favorite by them is use somebody  
 [PERSON 1:] Are you married? I will be married soon.  
 [PERSON 2:] haha, no time. I have got a novel to finish.  
 [PERSON 1:] What is your novel about?  
 [PERSON 2:] it is a love story. I really enjoy writing. do you write or read?  
 [PERSON 1:] Like to read. Spend most time watching football.  
 [PERSON 2:] yes I do I am and author for a living

Table 9: Example dialog between a human (Person 1) and the Key-Value Profile Memory Network with Self Persona.

Persona 1	Persona 2
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?

[PERSON 2:] I am doing good and you?

[PERSON 1:] I am well, just drinking some scotch and working on my writing

[PERSON 2:] I love to cook and bake but I rarely make time for it

[PERSON 1:] what do you usually bake?

[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?

[PERSON 1:] I used to be a juggler at a carnival but now I am an author

[PERSON 2:] I am a dancer, what do you do?

[PERSON 1:] I just told you. Do you have any pets?

[PERSON 2:] no, I am allergic to pets unfortunately.

[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!

[PERSON 2:] I have a cat. she has been with me through three wives, haha.

[PERSON 1:] Funny that you just told me you have no pets

[PERSON 2:] I do not have any animals.

Person\_1 is a human

Person\_2 is the bot  
(generative)

Table 7: Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).

# Meena: Generation is closing the gap

Adiwardana et al., 2020

- (Evolved) Transformers: 2.6B parameters
- Trained for 30 days on a TPU v3 pod (2,048 TPU cores !) on data from Reddit
- Paper [here](#)

Meena Conversation 7		Meena Conversation 1	
<b>Human:</b>	<i>Hi!</i>	<b>Human:</b>	<i>Hi!</i>
<b>Meena:</b>	<i>Hi! How's it going?</i>	<b>Meena:</b>	<i>Hey there! What's up?</i>
<b>Human:</b>	<i>doing okay, feeling a little bit under the weather</i>	<b>Human:</b>	<i>Same old, Monday blues...what are you up to?</i>
<b>Meena:</b>	<i>Aw, I hope you feel better soon!</i>	<b>Meena:</b>	<i>I'm procrastinating on a paper I'm supposed to be writing.</i>
<b>Human:</b>	<i>thanks! the weather is getting worse</i>	<b>Human:</b>	<i>are you thinking of pulling an all nighter?</i>
<b>Meena:</b>	<i>Oh no, I'm sorry. I hope it gets better for you soon.</i>	<b>Meena:</b>	<i>Possibly. What's your paper on?</i>
<b>Human:</b>	<i>no, it won't get better any time soon. we have to wait until spring</i>	<b>Human:</b>	<i>Nah, i just do the grunge work around here...one day I wish to write a paper that will change the world!</i>
<b>Meena:</b>	<i>Aww, sorry to hear that.</i>	<b>Meena:</b>	<i>What's the topic?</i>
<b>Human:</b>	<i>yea, I'm really feeling under the weather</i>	<b>Human:</b>	<i>not sure, any suggestions?</i>
<b>Meena:</b>	<i>Aw, well that's no good.</i>	<b>Meena:</b>	<i>Well what's the assignment on?</i>
<b>Human:</b>	<i>yea, any suggestion to get over feeling under the weather</i>	<b>Human:</b>	<i>breathing under water</i>
<b>Meena:</b>	<i>I wish I knew. Sorry I can't help.</i>	<b>Meena:</b>	<i>How to breathe underwater.</i>
<b>Human:</b>	<i>have you ever been feeling under the weather?</i>	<b>Human:</b>	<i>that's what i'm hoping to figure out!</i>
<b>Meena:</b>	<i>Not that I can remember, no.</i>	<b>Meena:</b>	<i>I hope you figure it out! Because I can't.</i>
<b>Human:</b>	<i>you must be a superman!</i>	<b>Human:</b>	<i>i'm sure you can if you try :) you're the best!</i>
<b>Meena:</b>	<i>Nah, I'm just lucky, I guess.</i>	<b>Meena:</b>	<i>Thank you, but I'm pretty sure you're the best.</i>
		<b>Human:</b>	<i>how was your weekend?</i>
		<b>Meena:</b>	<i>Pretty good. Yours?</i>
		<b>Human:</b>	<i>good, I practiced the violin</i>
		<b>Meena:</b>	<i>That's awesome! How long have you been playing?</i>
		<b>Human:</b>	<i>just started learning last month</i>
		<b>Meena:</b>	<i>That's awesome! What made you want to learn the violin?</i>

# Joint testing on different tasks?

- Dialog systems should be open-domain!
- Dialog is not monolithic:
  - Ask facts?
  - Ask for opinions (recommendations)?
  - Dialog combining facts and opinions?
  - General chit-chat (statements not questions)?
- Combination of all above in one end-to-end model?

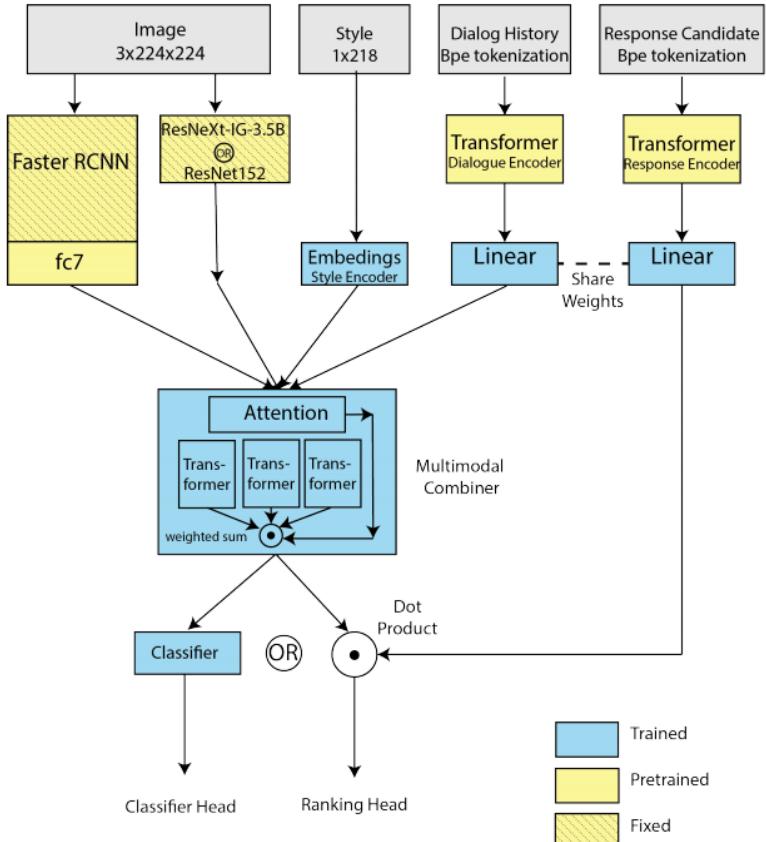
# Dialog Dodecathlon

Schuster et al., 2019

	Existing Approaches (independent)				MT + FT		All Tasks MT	
	Approach	PPL	Score	(Metric)	PPL	Score	PPL	Score
ConvAI2	(Lewis et al., 2019)	11.9	*20.7	F1	11.1	21.6	<b>10.8</b>	<b>21.7</b>
DailyDialog	(He et al., 2019)	11.1	-	F1	<b>10.4</b>	<b>18.2</b>	12.0	16.2
Wiz. of Wikipedia	(Dinan et al., 2018)	23.1	35.5	F1	<b>8.3</b>	<b>38.4</b>	8.4	<b>38.4</b>
Empathetic Dialog	(Rashkin et al., 2019)	21.2	6.27	Avg-BLEU	<b>11.4</b>	8.1	11.5	<b>8.4</b>
Cornell Movie	(He et al., 2019)	27.5	-	F1	<b>20.2</b>	<b>12.4</b>	22.2	11.9
LIGHT	(Urbanek et al., 2019)	*27.1	*13.9	F1	<b>18.9</b>	<b>16.2</b>	19.3	16.1
ELI5	(Lewis et al., 2019)	24.2	20.4	Avg-ROUGE-L	<b>21.0</b>	<b>21.3</b>	24.9	20.7
Ubuntu	(Luan et al., 2016)	46.8	-	F1	<b>17.1</b>	12.7	23.1	12.1
Twitter		-	-	F1	30.7	9.9	38.2	9.8
pushshift.io Reddit		-	-	F1	25.6	13.6	27.8	13.5
Image Chat	(Shuster et al., 2018b)	-	27.4	ROUGE-L ( $1^{st}$ turn)	<b>18.8</b>	<b>43.8</b>	22.3	39.7
IGC	(Mostafazadeh et al., 2017)	-	1.57	BLEU (responses)	11.9	<b>4.4</b>	12.0	4.3

# Multimodal chatbots

Ju et al., 2020



# Multimodal chatbots

Ju et al., 2020

	<b>Task</b> Coco	<b>TransResNet MMC</b> there is a broken tree log on the ground		<b>Task Context</b> VQA Does it appear to be rainy?	<b>TransResNet MMC</b> no
	<b>Task</b> Coco	<b>TransResNet MMC</b> A large grass covered field under a mountain.		<b>Task</b> Personality Captions (Style: Happy)	<b>TransResNet MMC</b> Wow what a beautiful and perfect shade of pink and red! I am so captivated!
	<b>Task</b> Flickr30k	<b>TransResNet MMC</b> A chaparral landscape scene void of human residence.		<b>Task</b> Personality Captions (Style: Attractive)	<b>TransResNet MMC</b> Wow I would love for someone to create this setting in the sand for me.
	<b>Task</b> Flickr30k	<b>TransResNet MMC</b> A plane flying sideways.		<b>Task</b> Image Chat (Style: Compassionate)	<b>TransResNet MMC</b> Round 1: Something about the pattern calms me. <b>TransResNet MMC</b> The architecture calms you.
	<b>Task Context</b> VQA What is the color of the mountain?	<b>TransResNet MMC</b> gray		<b>Task Context</b> Image Chat (Style: Emotional) Round 1: Airplanes are scary to get on, you never know if it will crash or not. Round 2: But these are professional pilots though.	<b>TransResNet MMC</b> They are, and for many people they mean a lot. My grandfather loved planes!

# Open topics

- Active learning from interactions
  - Reinforcement learning from interactions
  - Mixing information-retrieval with generative models
  - Multimodal still needs a lot of work
  - Open-domain: talking about anything, mixing goal-oriented and chit-chat.
- 
- Need interactions at scale!!

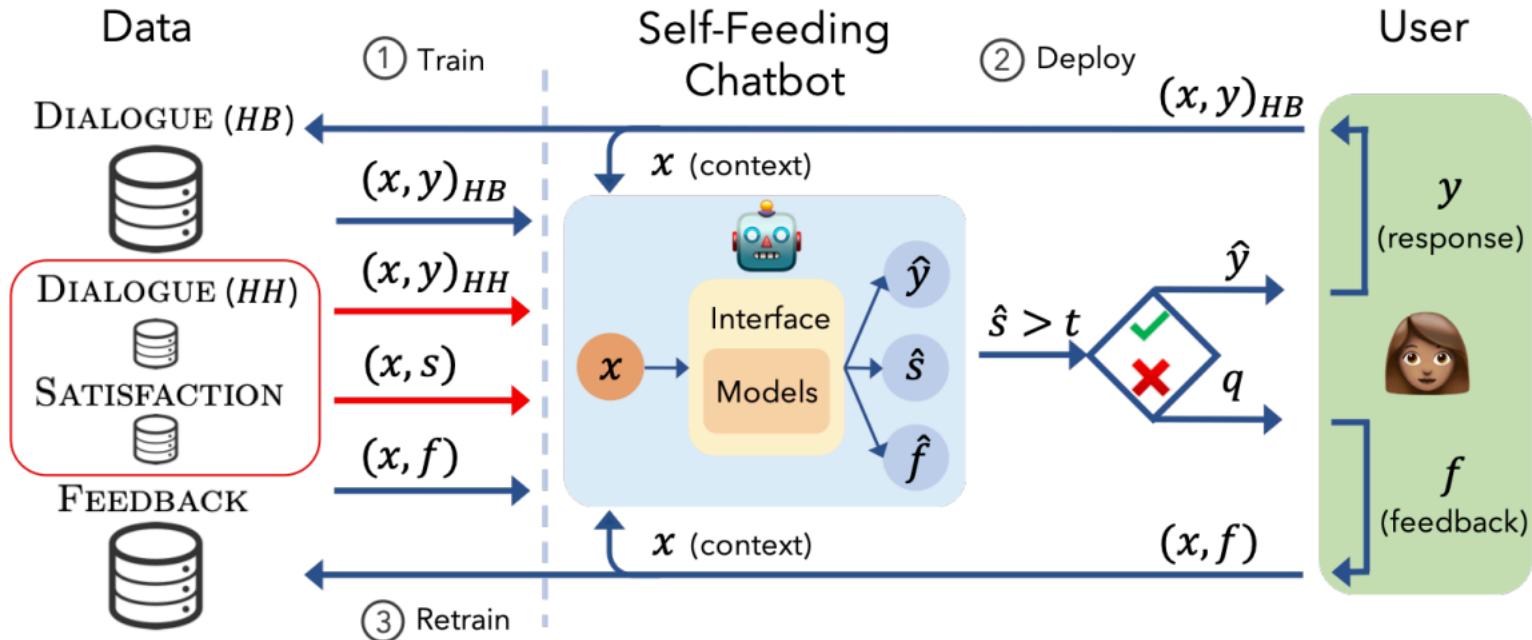
# Dialog is about interaction – can we use it?

Hancock et al., 2019



# A self-feeding chatbot

Hancock et al., 2019



# References Compositional Sequence Encoders

- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. NAACL.
- McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. NIPS.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. arXiv.
- Howard, J. & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. ACL.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. NIPS.
- Cheng, J., Dong, L., & Lapata, M. (2016). Long short-term memory-networks for machine reading. EMNLP.
- Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. ACL.
- Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. ICLR.
- Yang, Z., Zhao, J., Dhingra, B., He, K., Cohen, W. W., Salakhutdinov, R., & LeCun, Y. (2018). GLoMo: Unsupervisedly Learned Relational Graphs as Transferable Representations. arXiv.
- Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved semantic representations from tree-structured long short-term memory networks. ACL.
- Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016). Recurrent Neural Network Grammars. NAACL.

# References Interaction

- Cho, K., Gulcehre, B. V. M. C., Bahdanau, D., Schwenk, F. B. H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. EMNLP.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. NIPS.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. ICLR.
- Sukhbaatar, S., Weston, J., & Fergus, R. (2015). End-to-end memory networks. NIPS.
- Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... & Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. ICML.
- Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... & Badia, A. P. (2016). Hybrid computing using a neural network with dynamic external memory. Nature
- Grefenstette, E., Hermann, K. M., Suleyman, M., & Blunsom, P. (2015). NIPS.
- Henaff, M., Weston, J., Szlam, A., Bordes, A., & LeCun, Y. (2017). Tracking the world state with recurrent entity networks. ICLR.
- Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiský, T., & Blunsom, P. (2016). Reasoning about entailment with neural attention. ICLR.
- Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. ICLR.

# References

- Adversarial Examples for Evaluating Reading Comprehension Systems (Jia et al. 2017, EMNLP)
- Know What You Don't Know: Unanswerable Questions for SQuAD (Rajpurkar et al. 2018, ACL)
- Visual question answering: Datasets, algorithms, and future challenges (Kafle et al. 2017, Computer Vision and Image Understanding)
- Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (Goyal et al. 2017, CVPR)
- Reading Wikipedia to Answer Open-Domain Questions (Chen et al. 2017, ACL)
- Event2Mind: Commonsense Inference on Events, Intents, and Reactions (Rashkin et al. 2018, arXiv)
- Semantically Equivalent Adversarial Rules for Debugging NLP Models (Ribeiro 2018, ACL)
- Understanding Neural Networks through Representation Erasure (Li et al. 2016, arXiv)
- HotFLip: White-Box Adversarial Examples for NLP (Ebrahimi et al. 2017, arXiv)
- Anchors: High-Precision Model-Agnostic Explanations (Ribeiro et al. 2018, AAAI)
- Deep contextualized word representations (Peters et al. 2018, NAACL)
- Learned in Translation: Contextualized Word Vectors (McCann et al. 2017, NIPS)
- Supervised Learning of Universal Sentence Representations from Natural Language Inference Data (Conneau et al. 2017, EMNLP)
- Efficient Estimation of Word Representations in Vector Space (Mikolov et al. 2013, NIPS)
- Simple and Effective Semi-Supervised Question Answering (Dhingra et al. NAACL 2018)
- Neural Domain Adaptation for Biomedical Question Answering (Wiese et al. 2017, CoNLL)
- Improving Language Understanding by Generative Pre-Training (Radford et al. 2018, arXiv)
- Neural Skill Transfer from Supervised Language Tasks to Reading Comprehension (Mihaylov et al. 2017, arXiv)
- **Representing General Relational Knowledge in ConceptNet 5** (Speer and Havasi, LREC 2012)
- Learning to understand phrases by embedding the dictionary (Hill et al. 2016, TACL)
- Leveraging knowledge bases in lstms for improving machine reading (Yang et al. 2017, ACL)
- Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. (Mihaylov and Frank, 2018, ACL)
- Reading Wikipedia to Answer Open-Domain Questions (Chen et al. 2017, ACL)
- Evidence aggregation for answer re-ranking in open-domain question answering (Wang et al. ICLR 2018)
- Marco Baroni and Gemma Boleda: <https://www.cs.utexas.edu/~mooney/cs388/slides/dist-sem-intro-NLP-class-UT.pdf>
- News article: <https://www.independent.co.uk/infact/brexit-second-referendum-false-claims-eu-referendum-campaign-lies-fake-news-a8113381.html>

# References for Datasets

- Building a question answering test collection, *Voorhees & Tice SIGIR 2000*
- Besting the Quiz Master: Crowdsourcing Incremental Classification Games, *Boyd-Graber et al. EMNLP 2012*
- Semantic Parsing on Freebase from Question-Answer Pairs, *Berant et al. EMNLP 2013*
- Mctest: A challenge dataset for the open-domain macchine comprehension of text, *Richardson et al. EMNLP 2013*
- Teaching Machines to Read and Comprehend, *Hermann et al. NIPS 2015*
- WikiQA: A challenge dataset for open-domain question answering, *Yang et al. EMNLP 2015*
- Large-scale Simple Question Answering with Memory Networks, *Bordes et al. 2015 arXiv:1506.02075.*
- The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations, *Hill et al. ICLR 2016*
- SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Rajpurkar et al. EMNLP 2016*
- [SQuAD 2.0] Know What You Don't Know: Unanswerable Questions for SQuAD, *Rajpurkar and Jia et al. ACL 2018*
- Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks, *Weston et al. ICLR 2016*
- Constraint-Based Question Answering with Knowledge Graph, *Bao et al. COLING 2016*
- MovieQA: Understanding Stories in Movies through Question-Answering, *Tapawasi et al. CVPR 2016*
- Who did What: A Large-Scale Person-Centered Cloze Dataset, *Onishi et al. EMNLP 2016*
- MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, *Nguyen et al. NIPS 2016*
- The LAMBADA dataset: Word prediction requiring a broad discourse context, *Paperno et al. ACL 2016*
- WIKIREADING: A Novel Large-scale Language Understanding Task over Wikipedia, *Hewlett et al. ACL 2016*
- TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension, *Joshi et al. ACL 2017*
- Crowdsourcing Multiple Choice Science Questions, *Welbl et al. WNUT 2017*
- RACE: Large-scale ReAding Comprehension Dataset From Examinations, *Lai et al. EMNLP 2017*
- NewsQA: a Machine Comprehension Dataset, *Trischler et al. RepL4NLP 2017*
- Science Exam Datasets by the Allen Institute for Artificial Intelligence: <https://allenai.org/data/data-all.html>
- SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine, *Dunn et al. <https://arxiv.org/pdf/1704.05179.pdf>*
- Quasar: Datasets for Question Answering by Search and Reading. *Dhingra et al. 2017 <https://arxiv.org/abs/1707.03904>*
- Constructing Datasets for Multi-Hop Reading Comprehension across Documents, *Welbl et al. TACL 2018*
- The NarrativeQA Reading Comprehension Challenge, *Kočiský et al. TACL 2018*